DOCUMENT RESUME

ED 129 901 TM 005 743

AUTHOR Ryan, Joseph P.; Hamm, Debra W.
TITLE Practical Procedures for Increasing the Reliability of Classroom Tests by Using the Rasch Model.
PUB DATE [Apr 76]
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Factor Analysis; *Factor Structure; *Item Analysis; *Mathematical Models; *Multiple Choice Tests; Test Construction; *Test Reliability
IDENTIFIERS *Latent Trait Model; *Rasch Model; Test Theory

ABSTRACT

A procedure is described for increasing the reliability of tests after they have been given and for developing shorter but more reliable tests. Eight tests administered to 200 graduate students studying educational research are analyzed. The analysis considers the original tests, the items loading on the first factor of the test, and the items which fit the Rasch model. Adjustments for test length are considered. Tests shortened by deleting items which do not fit the Rasch model have a higher internal consistency reliability then the longer original tests. This finding contradicts the theorem of classical test theory which states that increasing test length increases test reliability. The explanation for the find is that the theorem of classical test theory assumes that all items are homogeneous. Items not fitting the Rasch model are not unidimensional in the "latent-trait" sense which corresponds to being non-homogeneous in the "classical-test-theory" sense. Deleting such items results in a proportionately longer test in that a larger proportion of the remaining items is homogeneous. The misfitting items would not have been deleted on the basis of a classical item analysis. (Author/BW)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

ED129901

# Practical Procedures for Increasing the Reliability of Classroom Tests by Using the Rasch Model

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Joseph P Ryan

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Joseph P. Ryan  Debra W. Hamm

College of Education

University of South Carolina

Paper Read at the Annual Meeting of the National Council on Measurement in Education, San Francisco, Calif., 1976

TM005 743

# Practical Procedures for Increasing Reliability of Classroom Tests by Using the Rasch Model

Joseph P. Ryan Debra W. Hamm
College of Education, University of South Carolina

## Introduction

This paper investigates the use of a Rasch latent trait model as a means of eliminating items from a teacher-made test in order to increase the internal consistency reliability of the test and to clarify the factor structure of the items. The latent trait model applied in this study is that proposed for the analysis of dichotomously scored data by Rasch, (1960, Chapters V, VI) and expanded upon by Wright (1967) and Wright and Panchapakesan (1969). The procedure described here contrasts the Rasch model to the procedures prescribed by classical test theory, which suggests that test reliability is increased by adding items to a test.

The classical perspective assumes that all items are parallel, have the same ratio of true score information relative to error score information, and that error score variance is randomly distributed with a mean of zero. From these assumptions one can easily demonstrate that increasing test length n times increases true score variation as the square of n and error variance only as n (Gulliksen, 1950; Magnusson, 1966; Lord and Novick, 1968). In short, classical test theory suggests that all items add true score information faster than they add error score information.

In contrast, the Rasch model argues that test items are not necessarily parallel and consequently some items more accurately manifest the latent trait being measured than other items. When some items do not measure the same trait, error information will differ among items which necessarily implies that at least some items will add error information faster than true score information.

Items which do not fit the Rasch model are those which add error information at a higher rate than the rest of the items on the test. Convenient algorithms for identifying such items are described by Wright and Panchapakesan (1969) and Wright and Mead (1975). Eliminating items that do not fit the Rasch model should increase the reliability of a test because it will delete non-parallel items, the items with greatest error variation. Magnusson suggests the reasonableness of this assertion when he writes:

> The internal consistency coefficient we obtain from KR20 will therefore be directly dependent on the correlations between the items in the test, i.e., on the extent to which the items measure the same variable. The more homogeneous the items are, the greater the numerical value of KR20 will be for a given number of items in the test (p. 119).

If deleting the items which do not fit the Rasch model makes the test items more homogeneous and increases the test reliability, some manifestation of these changes in the test should be apparent in the factor structure of the test items. The first factor should become more pronounced. In particular, the proportion of total test variation accounted for by the first factor should increase after the items which do not fit the Rasch model have been deleted because the remaining items will contain more true score information and will thus more accurately reflect the latent trait measured by the test.

Instruments and Sample

The results of eight teacher-made tests are used to examine the issues raised in the introduction. The tests were designed by the authors for a required graduate course in Educational Research Methods in the College of Education at the University of South Carolina. All tests consisted of four-option multiple choice items and were administered to two hundred students enrolled in the course during the Spring semester of 1975. Five of the

tests were formative tests (FT) used for diagnostic purposes and the other three were summative examinations (SE) used for grading purposes. The tests varied in length from 15 to 50 items.

Procedures

The reliability of the eight tests is analyzed in five ways. First, for each test the Kuder-Richardson Formula 20 index of reliability is calculated. Second, a Rasch item analysis is performed on each test using the Calfit program of Wright and Mead (1975). Items which do not fit the model, according to a chi-square test of fit, are eliminated from each test. For each test, revised by eliminating the items that did not fit the Rasch model, the KR20 index of reliability is calculated. Third, a principle components factor analysis is performed. The KR20 is then calculated for the test by considering only the items which load on the first factor.

A direct comparison of the KR20 calculated on the three versions of the eight tests described thus far could be misleading, since the tests in steps 2 and 3 differ in the number of items each contains and are shorter than the original test. Consequently, the reliability of these two tests are adjusted using the Spearman-Brown formula so that they are comparable to a test as long as the original test. Five KR20 reliability coefficients are calculated for each test. These are calculated for 1) all items on the original test, 2) the items which fit the Rasch model, 3) the items which fit the Rasch model corrected for the test length, 4) the items which load on the first factor, and 5) the items which load on the first factor corrected for the test length.

The effects on the factor structure of deleting the items which do not fit the Rasch model will be investigated by comparing the proportion of variance accounted for by the first factor using three sets of test items. For each of the eight tests, the analysis will be performed on 1) all items on the original test, 2) the items remaining on the test after the items which do not fit the Rasch model have been deleted, and 3) the items which load on the first factor of the initial factor analysis for all the items.

Results

The five KR20 coefficients for each of the eight tests are shown in Table 1. In parentheses to the right of each KR20 is the number of items on the test.

KR-20 RELIABILITY COEFFICIENTS

Table 1.

| Test | Original Test | Rasch Fitting Items | Rasch Fitting Items Adjusted For Test Length | First Factor Items | First Factor Items Adjusted For Test Length |
|---|---|---|---|---|---|
| FT 1 | .629 (15) | .622(14) | .638 | .467(4) | .766 |
| FT 2 | .808(30) | .798(23) | .837 | .766(11) | .899 |
| FT 3 | .737(20) | .769(17) | .796 | .700(11) | .809 |
| FT 4 | .741(20) | .792(18) | .808 | .688(9) | .831 |
| FT 5 | .528(15) | .697(13) | .726 | .767(6) | .893 |
| SE 1 | .890(50) | .910(46) | .916 | .738(10) | .934 |
| SE 2 | .809(30) | .858(27) | .870 | .835(19) | .889 |
| SE 3 | .798(40) | .830(37) | .841 | .461(11) | .757 |

The data in Table 1 shows that generally the highest reliability is achieved for the set of items that remain after deletion of items that do not fit the Rasch model. In all but two cases, FT 1 and FT 2, the test shortened by the removal of the Rasch misfitting items is more reliable than the longer original test. For FT 1 and FT 2 the difference between the reliability of the original test and the Rasch fitting test is almost negligible.

The reliability of the test composed only of items which load on the first factor is extremely high when one considers how few items there are on these tests. In one case (FT 5), the reliability of the first factor test is greater than the reliability of the original test and the Rasch constructed test. However, in all tests but FT 5, the reliability of the original test is greater than the reliability of the first-factor test.

When the tests are adjusted so that the reliabilities are comparable to tests equal in length to the original, tests formed by the items loading on the first factor are the most reliable. The tests composed of the items that fit the Rasch model adjusted for test length are more reliable than the original test, but not as reliable as the adjusted first-factor test.

The proportion of variation accounted for by the first factor of the three sets of items for the eight tests is shown in Table 2.

Proportion of Variation
Accounted for by the
First Factor

Table 2

| Test | Original Items | Rasch Fitting Items | First Factor Items |
|---|---|---|---|
| FT 1 | .17 | .18 | .47 |
| FT 2 | .18 | .18 | .36 |
| FT 3 | .11 | .20 | .28 |
| FT 4 | .20 | .21 | .33 |
| FT 5 | .24 | .29 | .61 |
| SE 1 | .19 | .21 | .39 |
| SE 2 | .21 | .23 | .30 |
| SE 3 | .15 | .15 | .38 |

It is not surprising to note that the test composed of items which load on the first factor of the original factor analysis have a well defined first factor. For all eight instruments, the first factor of these tests accounts for a higher proportion of variation than the first factor of the original tests and the first factor of the tests composed of items which fit the Rasch model. The first factor of the Rasch fitting items accounts for a slightly higher proportion of variation than the first factor on the original tests for each of the eight tests examined here.

Discussion and Implications

The preceding analyses generally support the suggestion that deleting test items which do not fit the Rasch model increases the reliability of the test and results in a more pronounced first factor. The magnitude of the improvements obtained by applying the Rasch procedure are not particularly striking but are notable because classical test theory suggests that reliability is increased by adding items to a test. The classical theory offers this suggestion based on the assumption that the additional items are parallel to the existing items. The validity of this strong assumption can be tested by the Rasch procedure. In a sense, the Rasch analysis produces a 'longer' test in that more of the items on a test are parallel to each other after the items which do not fit the model are deleted.

The results of the study also show that the items which load on the first factor of a test have the characteristics predicted by classical test theory. The items that load on the first factor are extremely homogeneous and show a high degree of reliability. It is important to point out, however, that items which load on the first factor do not necessarily measure a unidimensional trait. Subsequent Rasch analysis of the items which load on the first factors of the original tests indicates that the factors do not form unidimensional tests.

By way of qualification, it is important to point out that some of the items that did not fit the Rasch model would also have been deleted from the test on the basis of a classical item analysis. However, several items which did not fit in the Rasch analysis had very high item discriminations and clearly loaded on the first factor (loading on the first factor is operationally defined as a factor loading of .30 or better with no loadings on other factors

exceeding this value). These items would almost certainly be retained on a test analyzed by classical test theory procedures.

In summary, the procedure described in this paper offers practical advice to a teacher who has given a test and wishes to maximize its reliability. The teacher might be interested to know that adding additional parallel items to the test will theoretically increase its reliability the next time it is used. For a teacher who wishes to score students on a test that they have already taken, however, it is more useful to provide a procedure that can increase the reliability of the test by deleting items from the existing data set.

## References


Gulliksen, H., Theory of Mental Tests. New York: Wiley, 1950.

Lord, Frederic M. and Novick, Melvin R., Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley Publishing Company, 1968.

Magnusson, D., Test Theory. Reading, Mass.: Addison-Wesley Publishing Company, 1966.

Rasch, G., Probabilistic Models for Some Intelligence and Attainment Scores. Copenhagen: Danish Institute for Educational Research, 1960.

Willmott, Alan S., and Fowles, Diana E., The Objective Interpretation of Test Performance: The Rasch Model Applied. U.S. Distributor: Humanities Press Inc., Hillary House - Fernhill House, Atlantic Highlands, N.J., 1974.

Wright, B. D., "Sample-free Test Calibration and Person Measurement". Proceedings of the 1967 International Conference on Testing Problems. Princeton: ETS, 1968, pp. 85-101.

Wright, B. and Mead, R., "Calfit: Sample-free Item Calibration with a Rasch Measurement Model". Research Memorandum Number 18. Chicago: Department of Education, University of Chicago, 1975.

Wright, B.D., and Panchapakesan, N., "A Procedure for Sample-free Analysis": J. Educ. Psychol. Measmt., 24, pp. 85-101.